

Modélisation statistique cyclique des locations de Vélo'v à Lyon

Pierre BORGNAT, Patrice ABRY, Patrick FLANDRIN

¹CNRS, Université de Lyon, Laboratoire de Physique de l'ENS Lyon (UMR 5672 CNRS), Lyon, France

pierre.borgnat@ens-lyon.fr, patrice.abry@ens-lyon.fr, patrick.flandrin@ens-lyon.fr

Résumé – Vélo'v est le système urbain de location de vélos de Lyon. Son utilisation est analysée en s'attachant principalement au nombre de locations au cours du temps afin d'identifier les facteurs qui le déterminent. Après une analyse empirique qui révèle le caractère non stationnaire et cyclostationnaire (cycle sur la semaine) du nombre de locations horaires, nous formulons un modèle non stationnaire à l'échelle supérieure au jour et cyclostationnaire sur une période de la semaine. Une régression statistique linéaire permet de quantifier quels sont les facteurs pertinents expliquant la variation du nombre de locations. Nous montrons que ce nombre est contrôlé par la saison, le jour et l'heure, le temps qu'il fait et la taille (nombre d'abonnés) du système. Le modèle obtenu et son pouvoir prédictif sont alors confrontés aux données Vélo'v.

Abstract – Vélo'v is Lyon's community bicycle program. It is studied here at a global level, to assess the evolution with time of the number of hired bikes, and find the relevant factors to explain the evolution. An empirical analysis first reveals the daily and weekly patterns in a cyclostationary manner, jointly with the non-stationary evolutions over time-scales of the day and larger. Combining this model with linear statistical regression, a model is proposed for the prediction of the number of bikes hired per hour. We show that the season (its weather and the existence of holidays), the popularity of the program (as given by the number of subscribed users) and the time during the week are the most relevant factors to predict the number of locations at a given hour. This is confronted to the real database of Vélo'v trips.

1 Les Vélo'v à Lyon

Vélo'v est un système de location de vélos déployé dans le Grand Lyon depuis mai 2005 par JCDecaux. Il est proposé aux usagers de louer des vélos (3000 fin 2007, 4000 disponibles aujourd'hui), à retirer à l'une quelconque des 334 stations distribuées dans la ville et à reposer à n'importe quelle autre station [1]. Des équipements de ce type se généralisent dans un nombre croissant de grandes villes comme à Paris avec le système *Vélib'* depuis juillet 2007 (étudié par exemple dans [2]), ou *Bicing* lancé à Barcelone en mars 2007 [3]. Ces systèmes de location ouvrent deux types de questions scientifiques :

- des questions relatives à l'usage des transports, en offrant la possibilité d'analyses quantitatives sur les déplacements en vélo,
- des questions sur l'évolution et le fonctionnement du système telles que : quelle est l'activité du réseau ; voir si le service fonctionne correctement ; savoir si on peut mieux réguler la mise à disposition des vélos par exemple en prédisant si il est intéressant d'injecter parfois plus de vélos dans les stations.

L'étude proposée ici porte sur l'étude globale des locations de Vélo'v au fil du temps comme étape préliminaire à la compréhension de la dynamique et de la régulation du système. L'objectif est (i) de proposer une analyse empirique du nombre de locations, (ii) de formuler un modèle non stationnaire adapté aux données et (iii) l'utiliser pour prédire le nombre de locations à un jour et un horaire donné.

Les données anonymisées de trajets de locations Vélo'v, depuis la mise en place du système jusqu'à début 2008, nous ont

été fournies par JCDecaux et le Grand Lyon. D'autres études sont en cours pour étudier la répartition spatiale des trajets. Ici, nous nous concentrerons sur le comportement global. La quantité analysée est le nombre de vélos loués au fil du temps.

2 Étude du nombre de locations horaires

La première question est de choisir une échelle de temps adaptée pour reconstituer le nombre de locations en fonction de l'instant. Le compromis de ce choix de granularité est classique : il y a plus de fluctuations pour des petits temps d'agrégation Δ , mais augmenter Δ risque de lisser des évolutions pertinentes. La médiane des durées de location est de 11 minutes, et la durée moyenne d'une location est de 32 minutes. Le choix ici est de fixer $\Delta = 1h$ ce qui assure de gommer l'effet des trajets individuels en regardant des données agrégées à un temps plus long que leurs temps caractéristiques.

En Fig. 1 (a), on représente sur 2 ans (de décembre 2005 à décembre 2007) le nombre de locations horaires superposé à ce même nombre moyenné au jour et à la semaine. Deux propriétés ressortent. Premièrement, on constate un comportement non stationnaire de la moyenne qui évolue au cours du temps. L'interprétation est simple : le système n'est pas dans un régime stationnaire au début (au fil des mois le nombre d'utilisateurs, en particulier les abonnés, augmente et le nombre de stations existantes aussi) et l'utilisation des Vélo'v dépend du temps qu'il fait, donc de la saison. La deuxième caractéristique est une modulation forte sur une journée. Cette modulation a une période égale à une semaine qui s'interprète par le fait que

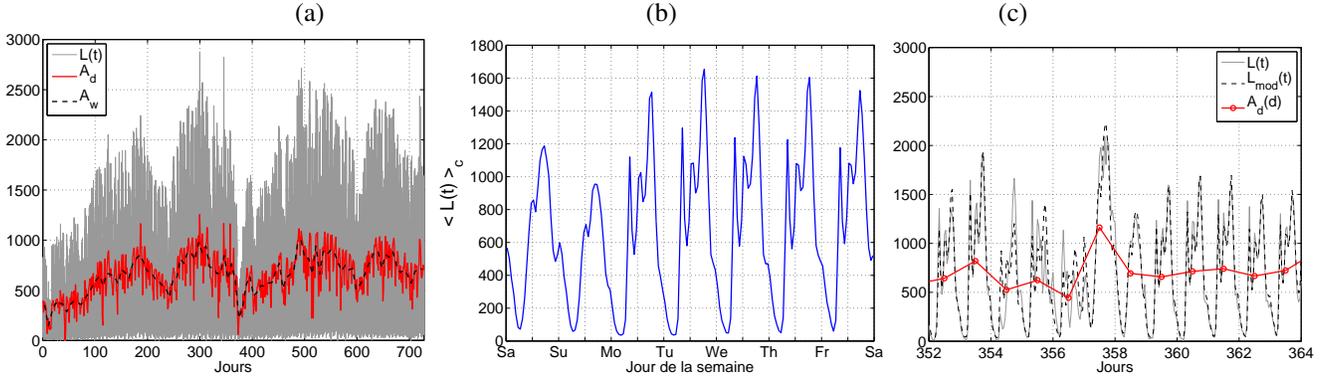


FIG. 1 – **Locations horaires.** (a) $L(t)$ tracé heure par heure, A_d moyenne sur la journée et A_w sur la semaine du 17.XII.2005 au 14.XII.2007. (b) Moyenne cyclique sur la semaine $\langle L(t) \rangle_c$ des locations horaires. (c) Détail autour du 356^e jour, le 8.XII.2006 (jour de la Fête des Lumières à Lyon) montrant que le modèle cyclique (voir éq. (3)) rend compte des évolutions usuelles, sauf quand une anomalie telle que le 8.XII apparaît dans la répétition cyclique.

les jours dans la semaine ne sont pas équivalents. En particulier, les samedis et dimanches on s’attend à des répartitions des utilisations des Vélo’v dans la journée qui changent par rapport aux jours ouvrables.

Ce sont ces doubles caractéristiques de non-stationnarité et de cyclostationnarité, périodique sur la semaine, que nous cherchons à modéliser.

3 Modèle temporel cyclique

Identifions tout d’abord l’évolution non stationnaire sur des échelles de temps supérieures à la période au jour d en faisant la moyenne sur un jour des locations horaires $L(t)$. On note $A_d(d)$ cette moyenne empirique sur la journée :

$$A_d(d) = \frac{1}{24} \sum_{(d)} L(t). \quad (1)$$

Nous puissions ensuite dans les méthodes cyclostationnaires [4] pour établir le comportement journalier. Une moyenne synchrone (cyclique) construit le gabarit hebdomadaire du nombre de locations à chaque heure. Elle est calculée heure par heure, aux instants $t = 0$ à $(24 \cdot 7 - 1) = 167$ h, en faisant la moyenne sur toutes les semaines disponibles dans les données :

$$\langle L(t) \rangle_c = \frac{1}{N_w} \sum_{d=0}^{N_w-1} L(t + 168 d), \quad (2)$$

où N_w est le nombre de semaines utilisées pour cette moyenne périodique. Cette répartition $\langle L(t) \rangle_c$ des locations dans la semaine est montrée en Fig. 1 (b). Il faut remarquer que dans la semaine ouvrable apparaissent 3 pics de locations (matin entre 8h et 9h, midi entre 12h et 14h et soir de 17h à 19h) correspondant aux 3 pointes usuelles de déplacements urbains. On note aussi que les samedis et dimanches ont des gabarits propres différents, où les pics sont moins marqués et où on voit principalement un pic large les après-midi. Ce comportement est assez proche de celui trouvé dans les données de Barcelone [3] (où l’étude porte sur le nombre de vélos en circulation).

On notera $A_{\text{mod}}(d_7) = \sum_{(d_7)} \langle L(t) \rangle_c$ le nombre total de locations pour chaque jour de la semaine d_7 (cette variable représente donc un jour entre le lundi et le dimanche). Nous proposons alors d’écrire le nombre de locations horaires par une modèle cyclique $L_{\text{mod}}(t)$ plus une fluctuation :

$$L(t) = L_{\text{mod}}(t) + F(t) = A_d(d) \frac{\langle L(t) \rangle_c}{A_{\text{mod}}(d_7)} + F(t), \quad (3)$$

où $F(t)$ est une fluctuation qui est le nombre de locations restantes qui n’est pas expliqué par ce modèle cyclostationnaire. Le modèle est illustré en Fig. 1 (c), en montrant son adéquation en général avec les données, ici sur deux semaines, bien qu’il échoue lors de conditions particulières comme dans cet exemple une fête usuelle le 8.XII à Lyon (mais parfois ce seront des conditions météorologiques inhabituelles, ce qui sera discuté plus loin).

4 Prédiction statistique et facteurs explicatifs du nombre de locations

Nous abordons la question suivante : pouvons-nous prévoir à partir des conditions extérieures au cycle temporel modélisé ci-dessus, le nombre de locations à un instant donné ? On dispose des données météorologiques (température et pluie), des indicateurs de déploiement du système que sont le nombre de vélos $N_v(d)$ mis à disposition dans le système et le nombre d’abonnés $N_s(d)$, donnés jour par jour. Enfin, les derniers facteurs généraux que nous avons sont les indications de certains jours où on s’attend à des comportements spécifiques différents (vacances et jours fériés ou grèves de transport). Le problème de prédiction est séparé en deux questions : la prédiction de l’amplitude non stationnaire $A_d(d)$ qui décrit combien de vélos sont loués un jour donné ; l’étude des fluctuations horaires $F(t)$ qui disent comment la donnée observée dévie du modèle de la journée. On se posera la question de savoir à quel point on peut prédire cette fluctuation heure par heure.

4.1 Prédiction de l'amplitude journalière $A_d(d)$.

Nous écrivons un modèle statistique de régression linéaire qui intègre les facteurs explicatifs suivants :

1. les conditions météorologiques, résumées par la température moyenne sur la journée $T(d)$, et le pluie tombée $R(d)$ (en mm de pluie) ; on prendra pour la température la déviation à sa moyenne $\delta T(d) = T(d) - \langle T(d) \rangle$ (en °C) ;
2. les indicateurs de déploiement et de popularité du programme : nombre de vélos disponibles $N_v(d)$, nombre d'abonnés $N_s(d)$; on prendra aussi les déviations $\delta N_v(d)$, $\delta N_s(d)$ par rapport aux dernières dates dans les données disponibles (décembre 2007) où le système a atteint un état proche de l'état actuel.
3. les jours spécifiques : $J_h(d)$ pour les vacances et jours fériés, $J_s(d)$ pour les grèves, où l'indicateur vaut 1 pour ces jours particuliers et 0 sinon.

On écrit le modèle de régression linéaire comme il suit :

$$\widehat{A}_d(d) = \alpha_0(d_7) + \alpha_1 \delta N_s(d) + \alpha_2 \delta N_v(d) + \alpha_3 \delta T(d) + \alpha_4 R(d) + \alpha_5 J_h(d) + \alpha_6 J_s(d), \quad (4)$$

Les indicateurs $\delta N_v(d)$, $\delta N_s(d)$, $\delta T(d)$ et $R(d)$ sont normalisés à une variance de 1 pour l'estimation des coefficients.

Terme constant $\alpha_0(d_7)$. Ce terme est d'abord ajusté avec le jour d_7 qui ne rend compte que de la position du jour dans la semaine (de lundi à dimanche) :

$$\alpha_0(d_7) = A_0 + c_1 (A_{\text{mod}}(d_7) - \langle A_{\text{mod}}(d_7) \rangle_{d_7}). \quad (5)$$

Cette dépendance est nécessaire puisqu'on a vu en Fig. 1 (b) que le nombre de locations attendu un jour donné dépend de sa position dans la semaine ; il est plus faible par exemple le week-end. Par minimisation de l'écart quadratique aux données, on obtient pour les deux constantes $A_0 = 17370 \pm 320$ (en nombre de Vélo'v) et $c_1 = 1.05 \pm 0.14$ (coefficient sans dimension). Les intervalles de confiance (IC) donnés sont ceux de la couverture à 95% sous hypothèse de loi gaussienne. Cette hypothèse sera discutée plus bas.

Le terme A_0 est caractéristique du nombre moyen de locations dans une journée (à l'état de référence), soit en moyenne 725 vélos loués par heure. Le terme correctif selon le jour est linéaire (c_1 est estimée proche de 1) qui prend en compte les différences d'un jour sur l'autre dans la semaine type.

Les autres facteurs. Ils sont obtenus par régression linéaire multi-variée (en minimisant l'écart quadratique aux données) et on indique dans la table 1 les valeurs estimées et les intervalles de confiance obtenus (à 95% de confiance et à nouveau sous hypothèse gaussienne). Le résultat du modèle $\widehat{A}_d(d)$ est comparé en Fig. 2 (a) aux amplitudes journalières réelles (sur la première année).

Les résultats nous éclairent sur la pertinence des facteurs explicatifs. Comme espéré, la température moyenne du jour joue en positif et la pluie en négatif sur le nombre de locations. Le

TAB. 1 – **Modèle linéaire pour $A_d(d)$, Éq. (4).** On donne les facteurs en jeu (et leur unité), leur valeur de référence ($N_s(d)$ et $N_v(d)$ fin décembre 2007, $\langle T(d) \rangle$; pour la pluie on donne $\langle R(d) \rangle$ même si la référence choisie est 0) et leur écart quadratique moyen (std.). En-dessous, on trouve les coefficients obtenus par régression linéaire multi-variée : la valeur estimée (est.) et les intervalles de confiance $[IC_-, IC_+]$ à 95 % (sous hypothèse gaussienne).

Facteur	$\delta N_s(d)$	$\delta N_v(d)$	$\delta T(d)$	$R(d)$	$J_h(d)$	$J_s(d)$
Unité	abonnés	vélos	°C	mm		
réf.	62 250	3 000	13.0	0.11		
std.	8 030	400	7.7	0.37		
coeff.	α_1	α_2	α_3	α_4	α_5	α_6
est.	1 860	-120	2270	-1280	-2900	20
IC ₋	1 210	-720	1980	-1520	-3700	-2900
IC ₊	2 560	+490	2560	-1030	-2100	+2900

nombre d'abonnés agit comme facteur positif mais, de manière étonnante, le nombre de vélos ne semble pas agir beaucoup ; la raison en est que les quantités N_s et N_v ont suivi une évolution similaire dans les 2 ans, rendant ces deux facteurs presque colinéaires (facteur de corrélation de 0.94). L'un des facteurs capture donc la majorité de la variation et l'autre le peu qui reste alors (avec un IC qui inclut 0). Enfin les vacances agissent, comme on pouvait le voir en Fig. 1 (a) pendant les étés (jours entre 200 et 250 la première année) ou les vacances d'hiver (autour de 370 à 390), en réduisant l'utilisation des vélos, avec autour de 3000 vélos de moins un jour de vacances comparé à un jour normal. Pour l'indicateur de facteur grève J_s , l'IC contenant 0 et étant très large, on se gardera de conclure : il n'y a tout simplement pas assez de jours de ce type recensés dans la base de données employée (moins de dix).

En Fig. 2 (b) on a calculé l'histogramme des erreurs de prédiction de l'amplitude $\widehat{A}_d(d) - A_d(d)$. La distribution des erreurs n'est pas gaussienne (ainsi que ne le sont pas tous les facteurs explicatifs) ; cependant, on constate que les erreurs ont une distribution plus concentrée qu'une gaussienne de même variance. Par conséquent, les IC donnés sont plutôt pessimistes et donnent tout de même une indication sur l'ordre de grandeur de l'incertitude réelle.

4.2 Étude fluctuations horaires $F(t)$

Passons à la prédiction horaire des locations. Le modèle a été écrit en éq. (3). Nous étudions les fluctuations $F(t)$ qui restent au-delà de l'amplitude $\widehat{A}_d(d)$ et du motif cyclostationnaire $\langle L(t) \rangle_c$. Leur déviation standard est de 210 (nombre de vélos loués par heure) environ (la moyenne étant nulle). Ces fluctuations sont corrélées d'heure à heure. Dans un premier temps, l'algorithme de Levinson permet d'estimer les paramètres d'un modèle AR pour $F(t)$: on trouve principalement un processus AR(1) avec un paramètre de l'ordre de 0.60 (et des dépendances à des temps plus longs sont inutiles).

On tente d'inclure dans ces déviations un autre facteur explicatif que l'analyse empirique suggère (voir aussi la Fig. 2

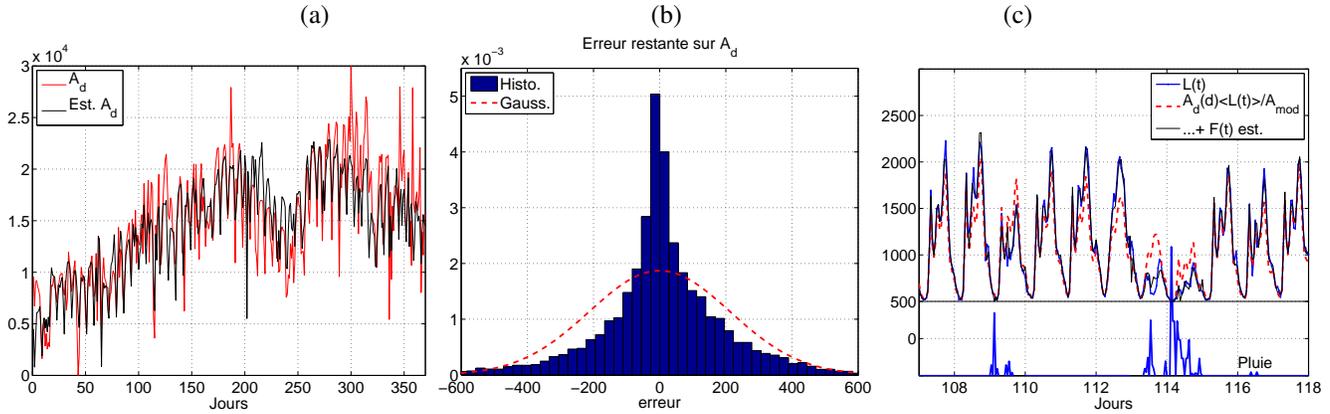


FIG. 2 – **Prédiction du nombre de locations horaires.** (a) Amplitude $\widehat{A}_d(d)$ sur la journée selon l'éq. (4). (b) Distribution des erreurs restantes sur $\widehat{A}_d(d)$ et comparaison à une loi normale de mêmes moyenne et variance. (c) Détail de la prédiction sans et avec la correction horaire $\widehat{F}(t)$, en superposition avec les données de pluie tombée ces jours-là.

(c) : la pluie tombée dans l'heure. Alors que la température a été mise principalement dans les effets non stationnaires (car ses variations principales sont saisonnières), les pluies sont des phénomènes souvent de plus courte durée et donc jouent beaucoup sur les fluctuations horaires. Nous étudions la régression statistique plus complète qu'un AR(1) :

$$F(t) = a_1 F(t-1) + \beta_1 R(t) + I(t), \quad (6)$$

où a_1 est un coefficient d'AR(1), β_1 le coefficient de régression linéaire associé à la pluie dans l'heure $R(t)$ (en mm) et $I(t)$ une innovation. À nouveau le critère d'adéquation est l'erreur quadratique. L'estimation (aux moindres carrés) donne : $a_1 = 0.59 \pm 0.02$ et $\beta_1 = -40 \pm 4$ (vélos/h / mm de pluie).

En combinant donc cette prédiction des fluctuations avec le modèle journalier, on arrive à réduire l'erreur quadratique moyenne pour l'estimation du nombre de locations à une heure donnée. La prédiction sans information horaire est :

$A_d(d) < L(t) >_c / A_{\text{mod}}(d_7)$ et a une erreur de 210 locations de vélos. Avec l'éq. (6), on préconise l'estimée $\widehat{F}(t) = a_1 F(t-1) + \beta_1 R(t)$ qui réduit cette erreur quadratique moyenne de prédiction à 104 locations horaires.

On montre en Fig. 2 (c) sur quelques jours le résultat de la prédiction du modèle et celui de la prédiction corrigée par heure avec les observations passées et la pluie. On voit l'amélioration apportée par la prise en compte des termes à l'heure, en particulier les jours de grande pluie.

5 Conclusion

On s'est intéressé à l'analyse et à la prédiction du nombre de locations de Vélo'v, répondant donc au problème de modélisation quantitative de ces données. Il faut noter que les techniques développées ici restent volontairement simples et visent dans un premier temps l'efficacité. Cependant, du fait des caractères non stationnaires et cyclostationnaires superposés, l'étude n'est pas aussi immédiate qu'il pourrait y paraître pour rendre

compte quantitativement des observations et ainsi comprendre quels sont les facteurs d'explication dominants.

- Une première conclusion de l'étude est de valider sur des données expérimentales une approche avec un modèle couplant une évolution temporelle non stationnaire $A_d(d)$ et un motif cyclostationnaire $< L(t) >_c$.
- En revenant aux questions sur les systèmes urbains de location de vélos, on conclue qu'un facteur dominant est le gabarit cyclique dans la semaine (dont l'origine est sociologique, avec les 3 pics de transports attendus chaque jour banalisé ouvrable).
- Enfin, on a mis en évidence une relation entre une évolution (non stationnaire) de la popularité du système et les facteurs de contrôle que sont le nombre d'abonnés (ou de vélos à disposition) et la saison—en particulier à travers les conditions météorologiques et les jours de vacances

Remerciements. Ce travail a été rendu possible grâce à l'aide de JCDecaux et du Grand Lyon et le soutien de l'IXXI (Institut des Systèmes Complexes de Lyon). Les auteurs remercient particulièrement Jean-Baptiste Rouquier ainsi que Céline Robardet, Antoine Scherrer, Pablo Jensen et Eric Fleury pour des discussions intéressantes.

Références

- [1] <http://www.velov.grandlyon.com/>
- [2] Girardin, F. "Revealing Paris Through Velib' Data" <http://liftlab.com/think/fabien/2008/02/27/revealing-paris-through-velib-data/> 2008.
- [3] Froehlich, J., Neumann, J., and Oliver, N. "Measuring the Pulse of the City through Shared Bicycle Programs" *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems - UrbanSense08*, Raleigh, North Carolina, USA, November 4, 2008.
- [4] Gardner, W., Napolitano, A., and Paura, L., "Cyclostationarity : Half a century of research", *Signal Processing* vol. 86 (4), p. 639–697, 2006.